Aurally Aided Visual Search Performance Comparing Virtual Audio Systems

Larsen, Camilla Horne; Lauritsen, David Skødt; Larsen, Jacob Junker; Pilgaard, Marc; Madsen, Jacob Boesen; Stenholt, Rasmus

Published in:
Spatial Cognition IX

Publication date:
2014

Document Version
Early version, also known as pre-print

Aurally Aided Visual Search Performance Comparing Virtual Audio Systems

Camilla H. Larsen, David S. Lauritsen, Jacob J. Larsen, Marc Pilgaard, Jacob B. Madsen & Rasmus Stenholt

School of Information and Communication Technology, Aalborg University, Denmark
{clarse10, dlauri10, jjla10, mpilga10}@student.aau.dk, {jbm, rs}@create.aau.dk

Abstract. Due to increased computational power, reproducing binaural hearing in real-time applications, through usage of head-related transfer functions (HRTFs), is now possible. This paper addresses the differences in aurally-aided visual search performance between a HRTF enhanced audio system (3D) and an amplitude panning audio system (panning) in a virtual environment. We present a performance study involving 33 participants locating aurally-aided visual targets placed at fixed positions, under different audio conditions. A varying amount of visual distractors were present, represented as black circles with white dots. The results indicate that 3D audio yields faster search latencies than panning audio, especially with larger amounts of distractors. The applications of this research could fit virtual environments such as video games or virtual simulations.

Keywords: Visual search, binaural audio, virtual environments, spatial audio, aurally aid, localization performance

1 Introduction

Spatial audio is an important feature in virtual environments as it helps users orient themselves, and can provide 360 aural awareness, independently of sight.

Previously, the possibility of more closely simulating binaural hearing in real-time was constrained by hardware performance. Even though typical hardware can handle the processing task today, still only few applications utilize this technology and instead a simple stereo amplitude panning method is used.

Prior to this study, we did an audio exclusive localization experiment, the results of which suggest that binaurally simulated sound (3D sound) improves localization performance, compared to amplitude-panned sound (panning sound) [1].

The comparison between 3D sound and panning sound in this study is exclusive to virtual environments and includes visual stimuli.

Our hypothesis is this: there is a significant difference in search latencies in aurally aided visual search, between using a panning- and a 3D audio system.

2 Experiment design

In this experiment, the participant had to locate a visual object within a virtual environment with changing audio- and visual conditions for each trial. The participant's task was to locate a visual target among a set of visual distractors, using a mouse to manipulate the virtual camera view. The experiment consisted of two independent variables: audio condition and amount of distractors. The audio condition consisted of three levels: no audio, panning audio and 3D audio. The amount of distractors consisted of eight levels: 0, 2, 4, 8, 16, 32, 64 or 128.

Before each trial a countdown appeared, counting from three to zero, which indicated when the participant could begin. When the participant had located the target, he was required to aim at the target with an on-screen cross-hair and perform a mouse click within the stimuli's bounds. If the participant hit the target, the system took control of the virtual camera and panned it to its starting position. The countdown appeared again, and the next trial began. Each participant went through 144 different trials, hence 48 trials for each audio condition, 6 for each amount of distractors. The order of audio rendering conditions followed a 3x3 Latin square design. The experiment was conducted as a within subject experiment.

The auditory stimuli used was 700 ms bursts of pink noise at a fixed audible level, with a silence period of 700 ms. The pink noise stimuli had 100 ms of linear fade, both in and out, leaving 500 ms at full intensity.

The visual stimuli used was a black circle with white dots in the middle. The black circle's visual angle was 4.7° of the virtual camera view. The target object was a black circle with an odd amount (one or three) of white dots inside, while the distractors had an even amount (two or four). The stimuli were also randomly rotated at 90° steps.

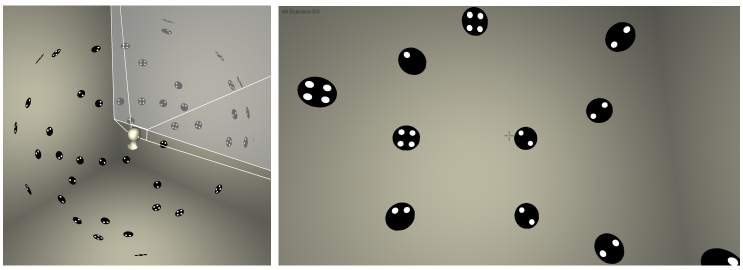

Fig. 1. Two depictions from the virtual environment. The image to the left shows a the setup of the environment with the virtual character in the center and 129 instances (note that the backfacing visuals cannot be seen) of visual stimuli distributed on the vertices of the icosahedron. The highlighted area represents the visual field. The image to the right is an example of how the screen could appear to a subject.

The positions of both the target and distractors were based on the vertices of a icosahedron, with 6 m in diameter. It was subdivided three times, resulting in 162 available points. The distractors and the target were placed at random discrete positions, each at one of the 162 available positions. There was only one target per trial. The amount of distractors for each trial was chosen through fixed randomization, assuring balance across all conditions. All stimuli were facing the participant's avatar at all time. The virtual settings was inside a box, the sides of which were dimly lit by yellow light. This was to provide participants with a sense of orientation. See figure 1 for a visual example of how one set might look. The participant's avatar was in the center of the icosahedron, with a distance of 3 m to every point.

3 Results and Discussion

A significant difference in search latencies was identified between the different audio conditions independent of distractor amount (Friedman test, $p < 0.001$) being $\bar{x} = 9.798$ seconds for the no audio condition, $\bar{x} = 3.588$ seconds for the 3D audio condition and $\bar{x} = 4.985$ seconds for the panning audio condition.

There was also a significant difference in search latencies with an increased amount of distractors (Friedman test, $p < 0.001$). Figure 2 illustrates the different combinations of audio conditions and distractor amounts and their significance level to each other. The linear increase in search latencies at increased distractor amounts suggests that the search task was not affected by preattentive search.

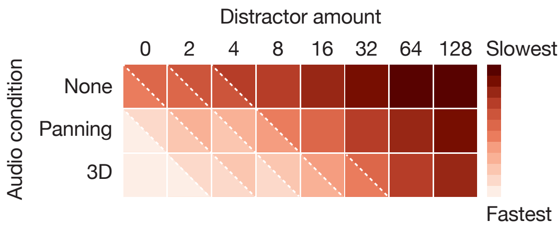

Fig. 2. The significance levels between combinations of audio conditions and amounts of distractors. White represents the combinations with the fastest search latency, red represents the slowest. Squares of different color are of a significant difference. Squares with two colors are of no significant difference to both identically colored whole squares.

Different starting positions for the target also created a significant difference (Friedman test, $p < 0.001$). If the target stimuli was placed at a vertical position exceeding 30° in the vertical plane relative the participant's initial look direction (making it not visible from the start) the participant would have to search through the vertical plane in order to reveal the target. These conditions increased the search latencies significantly (Friedman test, $p < 0.05$) with 0.802 seconds for the 3D audio condition and 2.194 seconds for the panning audio condition.

The accumulated time each participant had the target within their FOV, when fully visible, was also recorded. There was a significant difference (Friedman test, $p < 0.001$) in time, with mean values of $\bar{x} = 1.932$ seconds, $\bar{x} = 1.603$ seconds and $\bar{x} = 1.451$ seconds respectively for the no audio, panning audio and 3D audio conditions. This could be due to adaptive behavior, as the participant relies on audio for an early orientation cue. Typical orientation through binaural hearing is helped by moving the head and listening, potentially explaining the result.

The data indicates that there was a significant difference in search latencies between using the two different stimuli (Friedman test, $p = 0.003$) with mean latencies of $\bar{x} = 7.0$ seconds for the three dotted stimuli and $\bar{x} = 5.347$ seconds for the single dotted stimuli. The significant difference in latencies between the two visual target stimuli suggests that they were not equally susceptible to visual attention. Based on the works of Pomerantz and Cragin [3] we believe this is due to differences in emerging features. Both distractors and the three dot target stimuli contain more than two black dots, therefore emerging features such as proximity and orientations between the two or more points is present. The single dot stimuli can not have these features, except its positional feature from its single dot. This makes it stand out, and so, it becomes a Gestalt, leading it to emerge from the field of distractors. This can lead to faster acquisition times. We believe this did not have a great impact on our experiments results due to the within-subject design.

4 Conclusion

The results show that using 3D audio compared to panning audio decreases search latencies significantly (by 28%), confirming previous studies [2]. Furthermore, with increasing amount of visual distractors, 3D audio reduces search latencies compared to panning audio. This suggests that the visual distractors affect search performance, across all audio conditions. Search latencies were significantly decreased, by 9.5%, for 3D audio compared to panning audio, even with the visual stimuli being within the field of view, suggesting that auditory stimuli is used as an aid to visual search even within the field of view. The two visual stimuli gave significantly different search times, where the single-dot stimuli elicited faster search than the three-dot stimuli, which implies that our visual stimuli were not equal.

References

1. C. H. Larsen, D. S. Lauritsen, J. J. Larsen, M. Pilgaard, and J. B. Madsen. Differences in human audio localization performance between a HRTF- and a non-HRTF audio system. 2013.
2. J. P. McIntire, P. R. Havig, S. N. J. Watamaniuk, and R. H. Gilkey. Visual search performance with 3-d auditory cues: Effects of motion, target location, and practice. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(1):41–53, Feb. 2010.
3. J. Pomerantz and A. Cragin. Emergent features and feature combination. *Oxford Handbook of Perceptual Organization*, To be published.